



AntMiner+ Een Systeem van Kenniis-Ontginnende Mieren

RAF HAESSEN, DAVID MARTENS,
MANU DE BACKER, BART BAESENS

DE IMMENSE OPSLAGCAPACITEIT DIE
BEDRIJVEN VANDAAG DE DAG TER
BESCHIKKING HEBBEN, BEVAT DIKWILS
EEN KLUWEN VAN RUWE DATA WAARIN
WAARDEVOLLE KENNIS ZICH SCHUIL
HOUDT. DATA MINING OMVAT HET ONT-
GINNEN VAN DEZE VERBORGEN KENNIS
OP EEN GEAUTOMATISEERDE WIJZE.
VEELAL WORDT HIERBIJ ACCURAAUWHEID
ALS ENIGE PERFORMANTIE-INDICATOR
GEHANTEERD. DOCH BEGRIJPBAARHEID
KAN TEVENS EEN KRITISCHE VEREISTE
ZIJN IN DOMEINEN ZOALS MEDISCHE
DIAGNOSESTELLING EN CREDIT SCORING.
DE VOORGESTELDE TECHNIK,
ANTMINER+, IS GEÏNSPIREERD OP
HET GEDRAG VAN MIEREN EN RICHT ZICH
OP HET EXTRAHEREN VAN DERGELIJKE
ACCURATE ÉN BEGRIJPBARE MODELLEN.

VAN BIOLOGISCHE TOT ARTIFICIËLE MIERENSYSTEMEN

Individuele mieren zijn eenvoudige insecten met geli-
miteerd geheugen en vermogen om acties te onder-
nemen. Een mierenkolonie daarentegen vertoont zeer
complex gedrag en kan intelligente oplossingen vinden
voor problemen zoals het transporteren van grote voor-
werpen en het vinden van de kortste weg tussen een
voedselbron en het mierennest. Het intelligente gedrag,
“emergent behavior” genoemd, vloeit voort uit de zelf-
organisatie en indirecte communicatie tussen de mieren.
Mieren communiceren door middel van feromonen:
een mier laat deze chemische substantie achter op het
pad dat hij volgt. Omdat mieren in staat zijn de aanwe-

zigheid van deze substantie waar te nemen, ontstaat
indirecte communicatie met de andere mieren. Immers,
hoe hoger het feromoongehalte van het pad, des te
groter de kans dat de mier dit pad zal kiezen. Vermits de
mieren op het kortste pad meer feromonen per tijds-
eenheid kunnen afscheiden, zal na verloop van tijd
enkel dit pad versterkt worden en gekozen worden door
de mieren.

Laat ons dit verduidelijken met Figuur 1. Twee mieren
zijn op weg naar een voedselbron; initieel is er geen fero-
moon op de paden en is de kans dat de mieren het
bovenste of het onderste pad kiezen gelijk. De mier die
het onderste pad heeft gekozen is sneller terug bij het
nest, waardoor er dubbel zoveel feromonen op dit pad
liggen dan op het bovenste pad. Hierdoor is de kans dat
een volgende mier het onderste, kortere pad kiest dubbel
zo groot waardoor er nog meer mieren dit pad zullen
kiezen, zodat uiteindelijk (bijna) alle mieren het kortste
pad zullen volgen.



FIGUUR 1. PADSELECTIE BIJ MIEREN OP
BASIS VAN FEROMONEN

Artificiële mierenkolonies zijn gebaseerd op biologische
mierenkolonies en bestaan uit zeer simpele software
agenten, die net zoals mieren slechts een beperkt gehe-
ugen en vermogen hebben [3]. Een dergelijk systeem is
parallel en gedistribueerd, gezien de agenten in de pop-
ulatie simultaan en onafhankelijk van elkaar bewegen,
en dit zonder enige hogere instantie die bevelen geeft.
De agenten bewegen stochastisch op basis van twee
parameters, het feromoongehalte en een heuristische
waarde. Het feromoongehalte geeft aan hoeveel agenten
het spoor recentelijk hebben gevolgd, en de heuristische

waarde is een probleemafhankelijke parameter. Als een
agent aankomt op een beslissingspunt, heeft het pad
met het hoogste feromoongehalte en de grootste heuris-
tische waarde de meeste kans gekozen te worden. Eens
alle agenten aangekomen zijn in de eindbestemming,
wordt de omgeving aangepast: ten eerste daalt het fero-
moongehalte op alle paden door verdamping en ten
tweede wordt het pad van de beste agent versterkt door
het feromoongehalte te verhogen. Een nieuwe iteratie
van begin tot einde begint vervolgens. Heel dit proces
wordt herhaald tot de agenten convergeren naar een
(sub)optimale oplossing van het probleem.

DATA MINING: HET INFEREREN VAN KENNIS UIT EEN OVERVLOED VAN DATA

De groeiende populariteit van het Internet en de
recentelijk technologische vooruitgang op het vlak
van dataopslag hebben geleid tot een ware explosie van
ongestructureerde data. Het distilleren van bruikbare
patronen uit deze almaar groeiende stroom van ruwe
data vormt een bedrijfskritische uitdaging binnen onze
kenniseconomie. Data mining beoogt het extraheren
van dergelijke kennis op een geautomatiseerde wijze.
Classificatie is een onderdeel van data mining en omvat
het toekennen van datapunten aan een bepaalde klasse
op basis van specifieke kenmerken. Hiervoor wordt een
beroep gedaan op een hoeveelheid van beschikbare data
met gekende klasse. Classificatieproblemen doen zich
voor in tal van domeinen, zoals medische diagnosestel-
ling, fraudedetectie, credit scoring en het voorspellen
van aankoopprofielen in de marketing sector. Het resul-
taat van een classificatietechniek is een beslissingsmodel
waarmee toekomstige datapunten kunnen toegewezen
worden aan een bepaalde klasse.

(Vervolg op pag. 4)

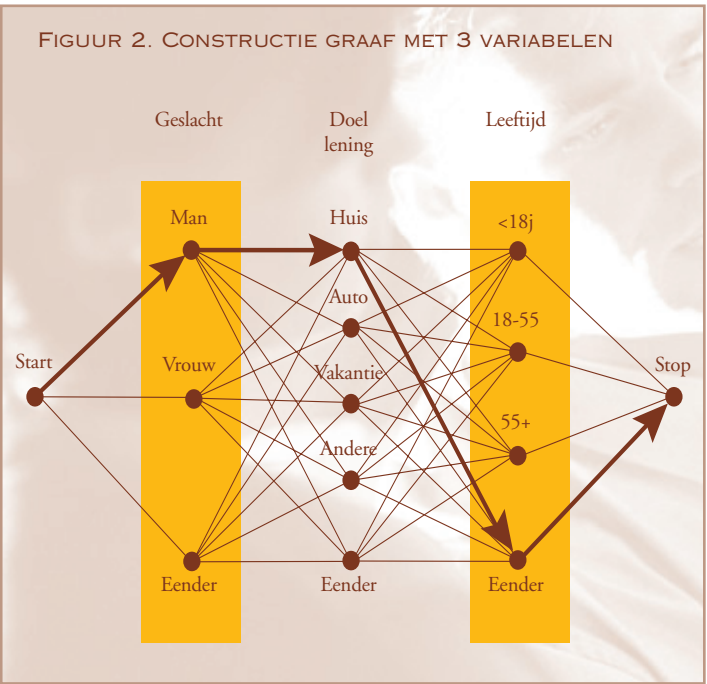
De performantie van een model wordt typisch gemeten door diens accuraatheid op een onafhankelijke dataset. Niet-lineaire classificatietechnieken, zoals neurale netwerken en support vector machines scoren het best volgens dit criterium [1], maar begrijpbaarheid van de gegenereerde modellen is ook een belangrijke vereiste. In de financiële sector bijvoorbeeld moeten kredietverleners een duidelijke reden kunnen voorleggen waarom krediet geweigerd wordt aan een klant, vage redenen zijn hierbij uit den boze! Onze aanpak houdt hiermee rekening en focust zich op het genereren van accurate én begrijpbare classificatiemodellen.

Het doel is om eenvoudige “als-dan-anders” regels af te leiden uit de data, waarbij het conditiegedeelte bestaat uit de conjunctie van een aantal termen. Elke term beschrijft ofwel een discrete waarde van een variabele, ofwel een continu interval van waarden. Daarom is elke term van de vorm “Variabele = Waarde”, ofwel “Variabele ≤ Waarde”, ofwel “Variabele ≥ Waarde”. Een voorbeeldregel bij kredietverlening is: “als *Inkomen ≥ €10.000 en gehuwd = ja dan kredietwaardig = ja anders kredietwaardig = nee*”.

DE LINK TUSSEN MIEREN EN DATA MINING IN ANTMINER+

In de voorgestelde aanpak wordt de oplossingsstrategie van een artificiële mierenkolonie toegepast op het classificatieprobleem [2]. De omgeving waarin de mieren bewegen, wordt voorgesteld door een gerichte, acyclische graaf zoals in Figuur 2. Voor elke variabele definiëren we een groep van knopen waarbij elke knoop een bepaalde waarde van die variabele voorstelt. Mieren die zich in een knoop van een bepaalde variabele bevinden, mogen enkel naar knopen van de volgende variabele gaan. Elke mier vertrekt in de startknoop met een lege regel en voegt in elke bezochte knoop een term aan de regel toe. Aangekomen in de eindknoop, beschrijft de mier een volledige regel.

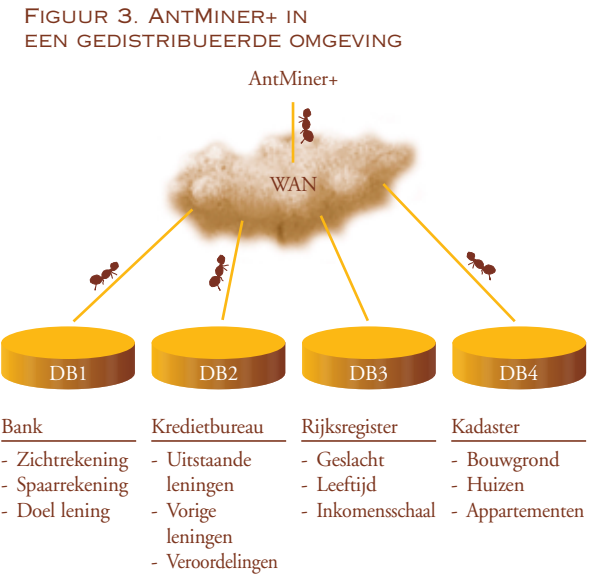
Om regels toe te laten die niet alle variabelen omvatten, wordt aan elke variabele een dummyknoop toegevoegd die eender welke waarde van de variabele aanneemt. Als een dergelijke knoop bezocht wordt, betekent dit dat de bijhorende variabele irrelevant is voor de classificatie en kan deze weggelaten worden uit de regel. Dit geeft aanleiding tot kortere en dus eenvoudigere regels. Het pad op de figuur dat vet staat aangeduid, beschrijft daarom de regel: “als *Geslacht = Man en Doel Lening = Huis dan Klant = Goed*”.



GEDISTRIBUEERDE DATA MINING MET ANTMINER+

Om het gebruik van AntMiner+ in een dynamische en gedistribueerde omgeving aan te tonen, hebben we een typische case uit de financiële wereld uitgewerkt, zie Figuur 2. De vraag of een klant al dan niet een lening mag aangaan, komt typisch neer op een binair classificatieprobleem om de betrouwbare betalers te onderscheiden van de onbetrouwbare. Stel dat een bank AntMiner+ wil gebruiken om een begrijpbaar model te construeren, waarmee beslissingen rond kredietwaardigheid kunnen verantwoord worden. De bank heeft allerlei klantgegevens ter beschikking zoals de saldo's van zichtrekeningen en spaarrekeningen. Ook heeft ze toegang tot data van een kredietbureau dat informatie van verschillende financiële instellingen consolideert. Andere bronnen van nuttige informatie zijn het rijksregister en het kadaster. Een gecentraliseerde aanpak zou heel wat dataverkeer over het netwerk veroorzaken en de verwerking zou moeizaam te realiseren zijn. AntMiner+ daarentegen stuurt mieren over het netwerk naar de relevante databases. De omgeving wordt dynamisch geconstrueerd over de verschillende sites, waarbij elke database een aantal variabelen bevat.

De gedecentraliseerde aanpak die hier voorgesteld wordt, maakt het mogelijk om op een snelle manier profielen van wanbetalers te ontdekken. Bovendien is de voorgestelde oplossing gemakkelijk schaalbaar en foutbestendig: mieren die verdwalen in het netwerk, of zelfs verbroken dataconnecties verlagen de performantie niet en AntMiner+ kan toch nog correct functioneren.



AntMiner+ werd toegepast op dergelijke datasets voor kredietvoorspelling. De geëxtraheerde classificatiemodellen zijn performant en begrijpbaar. Een voorbeeld van een set regels gegenereerd door AntMiner+ is te vinden in Tabel 1.

TABEL 1. “ALS-DAN-ANDERS” REGELS GEËXTRAHEERD DOOR ANTMINER+

als	(saldo zichtrekening < 100 EUR en duur > 15 M en krediet geschiedenis = geen en saldo spaarrekening < 500 EUR)
dan	klant = slecht
anders als	(doel = nieuwe auto/herstelling/onderwijs/ andere en krediet geschiedenis = geen/alle schulden afgelost in deze bank en saldo spaarrekening < 500 EUR)
dan	klant = slecht
anders als	(saldo zichtrekening < 0 EUR en doel = meubelen/huisraad/zaken en krediet geschiedenis = geen/alle schulden afgelost in deze bank en saldo spaarrekening < 250 EUR)

dan klant = slecht
anders als (saldo zichtrekening < 0 EUR en duur > 15 M en krediet geschiedenis = vertraging bij afbetalingen in het verleden en saldo spaarrekening < 250 EUR)
dan klant = slecht
anders klant = goed

CONCLUSIE

AntMiner+ is een techniek die de gedragsprincipes van mieren succesvol toepast op data mining. Niettegenstaande de beperkte mogelijkheden van een individuele mier, slaagt een AntMiner+ mierenkolonie erin om accurate, complexe, doch begrijpbare classificatiemodellen te ontginnen, die onmiddellijk inzetbaar zijn in het bedrijfsproces.

RAF HAESEN is als wetenschappelijk medewerker verbonden aan de KBC leerstoel in samenwerking met de Vlekho en de K.U.Leuven.
raf.haesens@econ.kuleuven.be

DAVID MARTENS is als wetenschappelijk medewerker verbonden aan het LIRIS-onderzoekscentrum in de Beleidsinformatica aan de Faculteit ETEW.
david.martens@econ.kuleuven.be

MANU DE BACKER is als wetenschappelijk medewerker verbonden aan het LIRIS-onderzoekscentrum in de Beleidsinformatica aan de Faculteit ETEW. Zijn onderzoek wordt gefinancierd door de Microsoft Research Chair.
manu.debacker@econ.kuleuven.be

BART BAESENS is deeltijds docent aan de Faculteit ETEW en lecturer (assistant professor) aan de School of Management, University of Southampton (United Kingdom).
bart.baesens@econ.kuleuven.be

REFERENTIES:

- [1] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.
- [2] M. De Backer, R. Haesen, D. Martens, and B. Baesens. A Stigmergy Based Approach to Data Mining, Proceedings of the Joint Australian Conference on Artificial Intelligence, *Lecture Notes in Computer Science*, accepted for publication, 2005.
- [3] M. Dorigo, V. Maniezzo, and A. Coloni. Positive feedback as a search strategy. Technical Report 91016, Dipartimento di Elettronica e Informatica, Politecnico di Milano, IT, 1991.